

# Admire LVQ—Adaptive Distance Measures in Relevance Learning Vector Quantization

Michael Biehl

Published online: 4 April 2012

© The Author(s) 2012. This article is published with open access at Springerlink.com

**Abstract** The extension of Learning Vector Quantization by Matrix Relevance Learning is presented and discussed. The basic concept, essential properties, and several modifications of the scheme are outlined. A particularly successful application in the context of tumor classification highlights the usefulness and interpretability of the method in practical contexts. The development and putting forward of Matrix Relevance Learning Vector Quantization was, to a large extent, pursued in the frame of the project *Adaptive Distance Measures in Relevance Learning Vector Quantization—Admire LVQ*, funded through the *Nederlandse Organisatie voor Wetenschappelijk Onderzoek* (NWO) under project code 612.066.620, from 2007 to 2011.

**Keywords** Machine learning · Adaptive distances · Prototype-based classification · Similarity-based clustering

## 1 Introduction

The comparison of observations with a set of reference data in terms of a similarity measure constitutes a natural and very intuitive approach to clustering or classification. The classic K Nearest Neighbor (KNN) scheme, arguably the most popular classifier, applies this concept with respect to previously stored, labeled example data.

Learning Vector Quantization (LVQ) as introduced by Kohonen [14] combines the idea with the representation of the reference data by a few prototypes. Their determination

in a training phase is often reminiscent of competitive learning [5]. Most frequently, a Nearest Prototype Classification (NPC) scheme is implemented in the working phase of LVQ classifiers.

Several improved versions of the basic LVQ algorithm have been suggested, including the cost function based Generalized LVQ (GLVQ) [15] and Robust Soft LVQ [20] which is guided by the maximization of likelihood ratios.

LVQ and other prototype-based classifiers share the attractive feature of being very intuitive and plausible, in contrast to many other learning systems. Prototypes are defined in the same space as the observed data and can be understood as typical representatives of their classes. This allows for straightforward interpretations of the classifier and facilitates discussions with application domain experts.

A key step in the design of such systems is the choice of a suitable measure of similarity or, rather, dissimilarity. For problems in which  $N$ -dimensional feature vectors represent the observations, standard Euclidean distance or other Minkowski metrics are employed most frequently without further justification. While such a choice appears plausible when all features are similar in nature, difficulties arise for feature vectors which comprise quantities of entirely different quality or order of magnitude. Standard distance measures are frequently sensitive to, for instance, rescaling or more general linear transformations of the features. Consequently, their naïve use is very often problematic in practice.

One very successful approach to overcome this difficulty is known as relevance learning in the literature [6, 12]. Its formulation is particularly transparent in the framework of prototype-based classifiers. The basic idea is to fix only the algebraic form of the dissimilarity in  $N$ -dim. feature space. Its parameters can then be optimized with respect to the available example data in the same training process as the prototypes. The concept is extremely versatile and allows for

---

M. Biehl (✉)  
Johann Bernoulli Institute for Mathematics and Computer  
Science, University of Groningen, P.O. Box 407,  
9700 AK Groningen, The Netherlands  
e-mail: [m.biehl@rug.nl](mailto:m.biehl@rug.nl)

the consideration of a variety of distances including generalized Euclidean metrics [6, 12] or less conventional choices like statistical divergences or measures tailored for the classification of functional data [13].

The main aim of the project *Adaptive Distance Measures in Relevance Learning Vector Quantization—Admire LVQ* was to develop and put forward an important extension of existing relevance learning schemes [7, 16]. The framework of Matrix Relevance LVQ (MRLVQ) was established which employs generalized Euclidean distances defined by quadratic forms, i.e. matrices of real-valued coefficients [9, 17–19]. In the project, a number of extensions and modifications were formulated and properties of the novel algorithms were studied theoretically and in computer experiments. Moreover, example applications were addressed, mainly in the context of biomedical data analysis.

In the next section, we exemplify the concept of MRLVQ in terms of the cost function based Generalized Matrix Relevance LVQ (GMLVQ) [17]. In Sect. 3 we highlight a recent medical application, which illustrates the usefulness and interpretability of the approach. We conclude with an outlook on current and future research in the context of adaptive distance measures in Sect. 4.

## 2 Matrix Relevance LVQ

An LVQ system for the classification of feature vectors  $\mathbf{x} \in \mathbb{R}^N$  is parameterized in terms of a number  $K$  of prototypes  $\{\mathbf{w}_j \in \mathbb{R}^N\}_{j=1}^K$  which serve as typical representatives of the classes  $c(\mathbf{w}_j) \in \{1, 2, \dots, C\}$ . Note that several prototypes may represent the same class.

Together with an appropriate dissimilarity measure  $d^\Lambda(\mathbf{w}, \mathbf{x})$ , the prototypes realize, for instance, an NPC scheme in which any possible input  $\mathbf{x}$  is assigned to the class of the closest prototype. In general, the superscript  $\Lambda$  may refer to a set of adjustable parameters. In the simplest version of Matrix Relevance LVQ, a global distance measure of the following quadratic form is employed [7, 16, 17]:

$$d^\Lambda(\mathbf{w}, \mathbf{x}) = (\mathbf{w} - \mathbf{x})^\top \Lambda (\mathbf{w} - \mathbf{x}) \quad (1)$$

which is specified by the  $(N \times N)$ -dim. relevance matrix  $\Lambda$ . In order to satisfy the minimal conditions  $d^\Lambda(\mathbf{x}, \mathbf{x}) = 0$  and  $d^\Lambda(\mathbf{x}, \mathbf{y}) \geq 0$  for  $\mathbf{x} \neq \mathbf{y}$ , a parameterization of the form  $\Lambda = \Omega^\top \Omega$  is assumed. Hence,  $\Omega$  defines a linear mapping of data and prototypes to a space in which standard Euclidean distance is applied:

$$d^\Lambda(\mathbf{w}, \mathbf{x}) = [\Omega(\mathbf{w} - \mathbf{x})]^2. \quad (2)$$

In its original formulation, no restrictions are imposed on the  $(N \times N)$ -matrix  $\Omega$  apart from a normalization  $\sum_j \Lambda_{jj} = \sum_{i,j} \Omega_{ij}^2 = 1$  [17]. The resulting NPC scheme implements general, piecewise linear boundaries which separate the classes.

The choice of prototype positions  $\{\mathbf{w}_j\}_{j=1}^K$  and elements  $\Omega_{ij}$  from a given set of example data  $\{\mathbf{x}^\mu, c(\mathbf{x}^\mu)\}_{\mu=1}^P$  with class labels  $c(\mathbf{x}^\mu)$  can be done according to iterative, heuristic procedures similar to the original LVQ algorithm [14]. Alternatively, the training process can be guided by suitable cost functions. A prominent example is the objective function of GLVQ suggested by Sato and Yamada [15]:

$$E = \sum_{\mu=1}^P \Phi[e(\mathbf{x}^\mu)] \quad \text{with } e(\mathbf{x}) = \frac{d^\Lambda(\mathbf{w}_J, \mathbf{x}) - d^\Lambda(\mathbf{w}_K, \mathbf{x})}{d^\Lambda(\mathbf{w}_J, \mathbf{x}) + d^\Lambda(\mathbf{w}_K, \mathbf{x})}. \quad (3)$$

Here, the index  $J$  identifies the *correct winner*, i.e. the closest prototype which carries the correct label,  $d^\Lambda(\mathbf{w}_J, \mathbf{x}) \leq d^\Lambda(\mathbf{w}_j, \mathbf{x})$  for all  $\mathbf{w}_j$  with  $c(\mathbf{w}_j) = c(\mathbf{x})$ . Correspondingly the *wrong winner*  $\mathbf{w}_K$  is the closest prototype that carries a label different from  $c(\mathbf{x})$ . The cost function is further specified by choosing the monotonic function  $\Phi$ , for instance as a sigmoidal or the identity in the simplest case. Note that the contribution  $e(\mathbf{x})$  is negative iff the feature vector  $\mathbf{x}$  is classified correctly by the given LVQ system. Hence, the minimization of  $E$  with respect to prototypes and the matrix  $\Omega$  can be interpreted as a large margin based training prescription. The actual optimization can be done by stochastic steepest descent as suggested in [17], where the full form of the corresponding gradient terms is given. Alternatively, batch gradient methods or more sophisticated optimization procedures can be applied.

A very important aspect of matrix relevance learning is its inherent tendency to yield low-rank matrices in the course of training: the resulting matrices  $\Lambda$  obtained by GMLVQ or heuristic LVQ updates become singular and are, generically, dominated by one or very few eigenvectors. Note that the distance measure (1) need not be strictly positive definite for the LVQ classifier to be meaningful. The convergence behavior can be understood by means of a mathematical analysis of the procedure under simplifying assumptions, see [2, 3]. As a consequence, the effective number of free parameters grows only linearly with  $N$ , while, nominally,  $N^2$  parameters have to be adapted. This explains why the introduction of relevance matrices in GMLVQ training has not led to overfitting effects in most practical applications or benchmark tests.

In the following, a few important variations of matrix relevance learning are mentioned, documenting the flexibility of the approach.

**Cost functions:** Generalized dissimilarities of the form (1) can be inserted in a multitude of distance based machine learning systems. In the context of LVQ, alternative cost functions can be employed, one example being the likelihood based objective of RSLVQ [20], see [18] for the incorporation of matrix relevances and the formulation of a likelihood-scheme replacing the simple NPC.

**Diagonal relevance matrices:** Restricting the measure (1) to diagonal matrices  $\Lambda$  and  $\Omega$  recovers the simpler and less powerful GRLVQ scheme [12]. Here, single dimensions in feature space are weighted or rescaled by the positive relevance factors  $\Lambda_{jj} = \Omega_{jj}^2$ .

**Local distance measures:** The flexibility of the matrix relevance approach can be greatly increased by using class-wise or local distance measures [17]. For instance, the assignment of a separate relevance matrix  $\Lambda^{(j)} = \Omega^{(j)\top} \Omega^{(j)}$  to each of the prototypes results in piecewise quadratic decision boundaries in the NPC scheme.

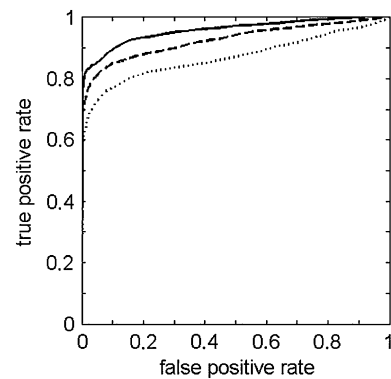
**Limited rank relevance matrices:** The basic formalism also allows to parameterize  $\Lambda$  in terms of a rectangular  $(N \times M)$ -matrix  $\Omega$  with  $M < N$ , where  $M$  explicitly limits the rank of  $\Lambda$ . Together with an appropriate penalty term added to the cost function, the system can be forced to exploit the maximum rank  $M$ . Therefore, explicit control of  $\text{rank}(\Lambda)$  can be achieved [7, 9, 19]. One motivation for imposing this restriction is to reduce the number of adaptive parameters explicitly. Despite the above discussed tendency of  $\Lambda$  to approach low rank, the optimization of  $(N \times N)$  matrices may become infeasible and costly in the presence of very high-dimensional data. Furthermore, the projections obtained by *Limited Rank Matrix LVQ* (LiRaM LVQ) facilitate the discriminative visualization of labeled data sets for  $M = 2$  or 3 [2, 7, 9].

### 3 Classification of Adrenal Tumors

Matrix relevance learning has proven useful in a number of benchmark problems and relevant practical applications from various contexts, including computer vision [10], bioinformatics [21], or content-based image retrieval [8]. In the following, a recent medical application is highlighted. Here, it is not the primary goal to report the results or technical aspects, which has been done in detail elsewhere [1, 4], but to emphasize the usefulness of the approach, in particular with respect to the interaction with domain experts.

The analysis of so-called *omics* data plays a role of increasing importance with respect to biomarker based diagnoses and personalized medicine. This comprises a multitude of methods based on genomics, proteomics, metabolomics or other patient specific data.

In collaboration with medical researchers from the *European Network for the Study of Adrenal Tumors* ([www.ensat.org](http://www.ensat.org)) and, in particular, from the Medical School of the University of Birmingham, UK, a diagnosis tool was developed for the detection of malignancy in adrenal tumors [1]. Matrix relevance learning was applied in the analysis of steroid metabolomics data, representing the 24 h excretion of 32 steroid biomarkers in patients with adrenal tumors. In a first



**Fig. 1** Receiver Operating Characteristics with respect to the detection of adrenocortical carcinoma in patients with adrenal tumor. The curves correspond to, from bottom to top, GLVQ using Euclidean distance in the 32-dim. feature space (dotted line), GRLVQ employing diagonal relevances only (dashed), and GMLVQ with a global  $(32 \times 32)$ -dim. relevance matrix (solid). See [1, 4] for the precise set up of the validation experiments

retrospective study the target was to assign each steroid profile to the class of benign adrenocortical adenoma, ACA for short, or malignant adrenocortical carcinoma (ACC).

By means of standard validation procedures, we demonstrated that the resulting classifier achieves very good sensitivity (true positive rate) and specificity (1-false positive rate) with respect to the detection of malignant ACC. Figure 1 displays the obtained Receiver Operating Characteristics [11] with respect to validation data. It shows that relevance learning and, in particular, matrix relevances clearly improve the performance over naive Euclidean distance in the framework of GLVQ. Standard classifiers and statistical methods of similar complexity achieved inferior performance and displayed strong overfitting effects; this included the Support Vector Machine with quadratic kernel, Fisher Linear Discriminant Analysis, and Logistic Regression, see [1, 4] for details.

Moreover, GMLVQ facilitated novel insights into the problem and provided an excellent basis for discussions in this interdisciplinary project. Besides interpretable prototypes for ACA and ACC steroid profiles, the method delivered the corresponding global relevance matrix which quantifies the significance of single steroid markers and combinations thereof. A number of markers were identified as highly relevant for the GMLVQ-based classification, while considerations on the level of univariate statistics showed no significant correlation with the class memberships [1]. This reflects the strength of matrix relevance learning as a truly multivariate, interpretable approach.

Based on the obtained matrix relevances, a panel of nine particularly relevant steroid markers was selected. Clearly, the achievable classification is limited compared to the use of the full panel of 32 markers. However, our analysis showed that the performance in terms of the ROC is only

slightly inferior [1, 4]. Limiting the analysis to a smaller number of markers facilitates an efficient technical realization of the test as a promising, non-invasive diagnosis tool [1]. Prospective studies concerning novel patient data will be required to substantiate this claim. Further aims of this ongoing line of research include the consideration of larger pools of potential markers, the incorporation of additional clinical data, and the monitoring of patients during and after treatment. Moreover, the potential identification of subtypes of ACA and ACC tumors should become feasible as larger amounts of patient data become available.

#### 4 Summary and Outlook

Within the NWO funded project *Admire LVQ*, Matrix Relevance LVQ has been established as a versatile and powerful framework for the classification of multidimensional data [7, 16]. The basic concept was developed and properties of various MRLVQ based algorithms were studied systematically. Benchmark tests and practical applications, mainly from the biomedical domain, showed that MRLVQ constitutes a competitive classification tool. Moreover, the emerging plausible systems provide valuable insights into the problems at hand and facilitate fruitful interdisciplinary collaboration.

Forthcoming studies will address important modifications of the basic MRLVQ scheme. For instance, block or triangular matrices could be used to reflect prior domain knowledge about the dependence or interaction of features. Along the same lines, specific forms of matrix relevance learning can be designed for the classification of functional data where *neighboring* features may be highly correlated, see [13] for first ideas concerning diagonal relevance matrices. Heterogeneous data sets comprising features from various sources and of different nature, for instance combinations of numerical and categorical data, require the design of modified adaptive distance measures and improved prototype schemes. Imposing suitable sparsity constraints on the relevance matrices opens new routes to feature selection which are currently under investigation. Moreover, the many attractive features of MRLVQ will be taken advantage of in further relevant and up to date practical applications.

**Acknowledgements** This project would have been impossible without the groundbreaking ideas and contributions of numerous colleagues. Special thanks go to Wiebke Arlt, Kerstin Bunte, Barbara Hammer, and Petra Schneider.

This work was supported by the *Nederlandse Organisatie voor Wetenschappelijk Onderzoek* (NWO) under project code 612.066.620.

**Open Access** This article is distributed under the terms of the Creative Commons Attribution License which permits any use, distribution, and reproduction in any medium, provided the original author(s) and the source are credited.

#### References

1. Arlt W, Biehl M, Taylor A, et al (2011) Urine steroid metabolomics as a biomarker tool for detecting malignancy in adrenal tumors. *J Clin Endocrinol Metab* 96:3775–3784
2. Biehl M, Bunte K, Schleif FM, Schneider P, Villmann T (2012) Large margin linear discriminative visualization by matrix relevance learning. In: *Proc international joint conference on neural networks (IJCNN 2012)*. (in press)
3. Biehl M, Hammer B, Schleif FM, Schneider P, Villmann T (2009) Stationarity of relevance matrix learning vector quantization. *Tech rep MLR 01/2009*, University of Leipzig, Machine Learning Reports
4. Biehl M, Schneider P, Smith D, et al (2012) Matrix relevance LVQ in steroid metabolomics based classification of adrenal tumors. In: Verleysen M. (ed) 20th european symposium on artificial neural networks (ESANN 2012), pp 423–428. d-side publishing
5. Bishop CM (2007) *Pattern recognition and machine learning*. Springer, Berlin
6. Bojer T, Hammer B, Schunk D, von Toschanowitz KT (2001) Relevance determination in learning vector quantization. In: Verleysen M (ed) *Proc of Europ symp on art neural networks (ESANN)*, pp 271–276. d-side
7. Bunte K (2011) Adaptive dissimilarity measures dimension reduction and visualization. PhD thesis, University of Groningen
8. Bunte K, Biehl M, Jonkman M, Petkov N (2011) Learning effective color features for content based image retrieval in dermatology. *Pattern Recognit* 44:1892–1902
9. Bunte K, Schneider P, Hammer B, Schleif FM, Villmann T, Biehl M (2012) Limited rank matrix learning, discriminative dimension reduction and visualization. *Neural Netw* 26:159–173
10. Denecke A, Wersing H, Steil J, Körner E (2009) On-line figure-ground segmentation with adaptive metrics in generalized LVQ. *Neurocomputing* 72:1470–1482
11. Fawcett T (2006) An introduction to ROC analysis. *Pattern Recognit Lett* 27:861–874
12. Hammer B, Villmann T (2002) Generalized relevance learning vector quantization. *Neural Netw* 15(8–9):1059–1068
13. Kästner M, Hammer B, Biehl M, Villmann T (2012) Functional relevance learning in generalized learning vector quantization. *Neurocomputing* (in press)
14. Kohonen T (1997) *Self-organizing maps*, 2nd edn. Springer, Berlin
15. Sato A, Yamada K (1996) Generalized learning vector quantization. In: Touretzky DS, Mozer MC, Hasselmo ME (eds) *Advances in neural information processing systems* 8. MIT Press, Cambridge, pp 423–429
16. Schneider P (2010) Advanced methods for prototype-based classification. PhD thesis, University of Groningen
17. Schneider P, Biehl M, Hammer B (2009) Adaptive relevance matrices in learning vector quantization. *Neural Comput* 21:3532–3561
18. Schneider P, Biehl M, Hammer B (2009) Distance learning in discriminative vector quantization. *Neural Comput* 21(10):2942–2969
19. Schneider P, Bunte K, Stiekema H, Hammer B, Villmann T, Biehl M (2010) Regularization in matrix relevance learning. *IEEE Trans Neural Netw* 21:831–840
20. Seo S, Obermayer K (2003) Soft learning vector quantization. *Neural Comput* 15(7):1589–1604
21. Strickert M, Schneider P, Keilwagen J, Villmann T, Hammer B (2008) Discriminatory data mapping by matrix-based supervised learning metrics. In: *Proc third intl workshop on artificial neural networks in pattern recognition*. Springer, Berlin, pp 78–89



**Michael Biehl** is Associate Professor with tenure at the Johann Bernoulli Institute for Mathematics and Computer Science, University of Groningen, The Netherlands. He received a Ph.D. in Physics from the University of Gießen in 1992 and the *Habilitation* in Theoretical Physics from the University of Würzburg in 1996. His main research interest is in the theory and application of machine learning techniques with recent emphasis on biomedical applications.